ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Διπλωματική Εργασία:

Δημιουργία Συστήματος Συστάσεων Βασισμένο σε Χωροχρονικές Πληροφορίες

Λιάπατας Γεώργιος, Μ1271

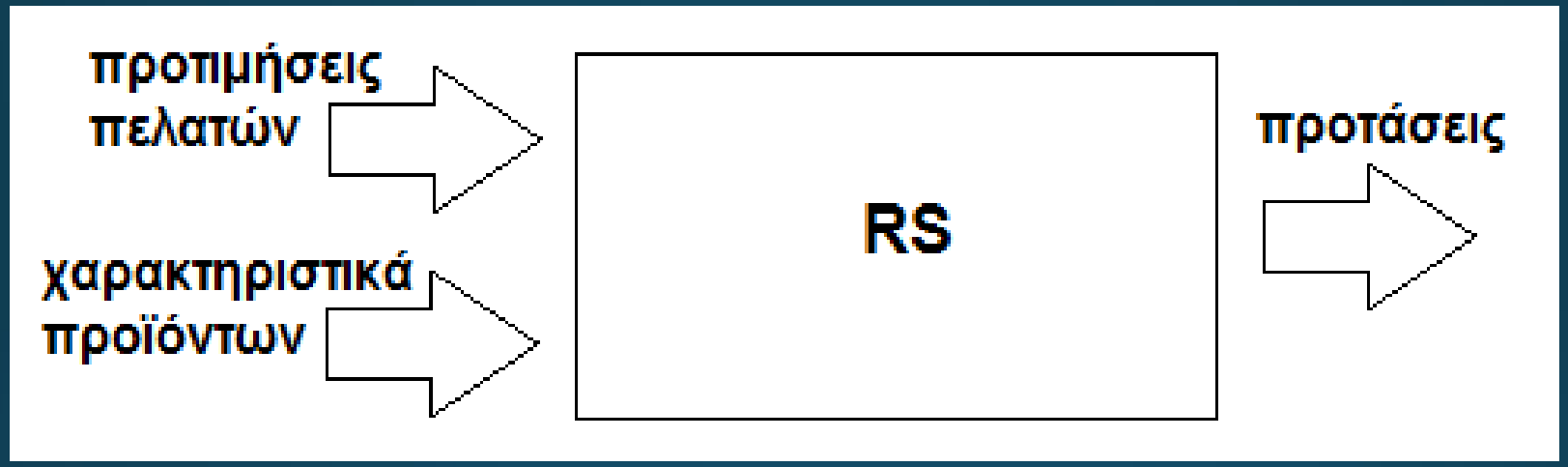
Επιβλέποντες: **Ευστάθιος Χατζιευθυμιάδης, Αναπληρωτής Καθηγητής ΕΚΠΑ**
Κωσταντίνος Κολομβάτσος, Διδάκτωρ ΕΚΠΑ

Περιεχόμενα

1. Εισαγωγή

- Τι είναι Σύστημα Συστάσεων;
  - Οντότητα που εξάγει εξατομικευμένες συστάσεις προς τους χρήστες
- Τι προσφέρει;
  1. Μετατροπή περιηγητή σε αγοραστή
  2. Αύξηση της εμπιστοσύνης
  3. Αύξηση παράλληλων πωλήσεων
  4. Αύξηση των πωλήσεων υπεραξίας
- Λειτουργεί όπως όλα τα συστήματα (είσοδος, επεξεργασία, έξοδος)

1.Εισαγωγή-Συνέχεια



- Είσοδος:
  - Προτιμήσεις πελατών
  - Χαρακτηριστικά προϊόντων
  - Συσχετίσεις
- Έξοδος
  - Συστάσεις

Είσοδος-Έξοδος

- Είσοδος:
  - Προέλευση:
    1. Σχετικά με τον πελάτη
       - Τάση: δεδομένα από περιήγηση του πελάτη στην ιστοσελίδα
    2. Σχετικά με μία κοινότητα πελατών
  - Διαδικασία λήψης δεδομένων:
    1. Διαφανής διαδικασία
       - Παράδειγμα διαφανούς εισόδου: καταγραφή πλοήγησης
    2. Είσοδος προκαλούμενη από το χρήστη
       - Προκαλούμενη από το χρήστη: βαθμολογία, ερωτηματολόγιο
- Έξοδος:
  - Εξαρτάται από την ποιότητα και την ποσότητα της εισόδου
  - Μορφή: Προβλέψεις βαθμολογίας, συστάσεις ή προτροπή προς το χρήστη για δοκιμή προϊόντος

Κεφάλαιο 2

Συστήματα Συστάσεων

2. Συστήματα Συστάσεων

- Κύριες Κατηγορίες Συστημάτων Συστάσεων
  1. Βασισμένα στο περιεχόμενο
  2. Συνεργατικού φιλτραρίσματος
  3. Βασισμένα στη γνώση
  4. Βασισμένα σε Δημογραφικά χαρακτηριστικά
  5. Υβριδικά

1) Βασισμένα στο Περιεχόμενο

- Προτάσεις προϊόντων όμοιων με αυτών που έχει προτιμήσει ο χρήστης στο παρελθόν
- Προφίλ για κάθε χρήστη
- Απαιτεί αξιολόγηση από το χρήστη
- Μηχανισμός ανάκτησης των χαρακτηριστικών που περιγράφουν ένα προϊόν
- Διαδικασία παραγωγής συστάσεων:
  1. Ανάλυση χαρακτηριστικών προϊόντος
  2. Σύγκριση με το προφίλ του χρήστη
  3. Απόφαση για σύσταση ή μη του προϊόντος

1)Βασισμένα στο Περιεχόμενο

- Προφίλ χρήστη:
  - Πληροφορίες για ενδιαφέροντα-προτιμήσεις του χρήστη
  - $W_u=(W_{u1}, W_{u2}, \ldots, W_{un})$, W: βάρη χαρακτηριστικών χρήστη
- Προφίλ αντικειμένου
  - Χαρακτηριστικά προϊόντος
  - Παράδειγμα ταινίας: είδος, σκηνοθέτης, ηθοποιοί, έτος κυκλοφορίας
  - $W_i=(W_{i1}, W_{i2}, \ldots, W_{in})$, W: βάρη χαρακτηριστικών χρήστη
- Πρόβλεψη σύστασης: συνάρτηση χρησιμότητας U(c, s), όπου c χρήστης και s αντικείμενο

2)Συνεργατικό Φιλτράρισμα

- Προτείνονται προϊόντα που αρέσουν σε παρόμοιους χρήστες
- Παρόμοιοι χρήστες: «γείτονες»
- Οι γείτονες έχουν ισχυρό συσχετισμό
- Παράδειγμα: (πρόβλεψη βαθμολογίας)

	Προϊόν 1	Προϊόν 2	Προϊόν 3	Προϊόν 4	Προϊόν 5	Προϊόν 6
Χρήστης Α	4	5		2		3
Χρήστης Β	3	**5**		**2**		**3**

2)Συνεργατικό Φιλτράρισμα

- Διαφορά με συστήματα βασισμένα στο περιεχόμενο:
  1. Ασχολείται με τις εκτιμήσεις και **όχι** με το περιεχόμενο
  2. Κάνει προτάσεις διαφορετικών προϊόντων από τις προηγούμενες εκτιμήσεις του χρήστη
- Προβλήματα μεθόδου:
  1. Ψυχρή εκκίνηση
  2. Αραιή πυκνότητα εκτιμήσεων
  3. Ασυνήθιστες προτιμήσεις χρηστών
  4. Δεν ασχολείται με πληροφορίες περιεχομένου
- Χρησιμοποιείται συνδυασμός συστημάτων βασισμένων στο περιεχομένο και συνεργατικού φιλτραρίσματος

3)Βασισμένα στη Γνώση

- Προτείνουν αντικείμενα που εντάσσονται σε συγκεκριμένο πεδίο γνώσης
- Χρήσιμα όταν δε μπορεί να εφαρμοστεί συνεργατικό φιλτράρισμα ή σύστημα βασισμένο στο περιεχόμενο
- Παράδειγμα: διαμερίσματα, αυτοκίνητα, υπολογιστές, χρηματοπιστωτικές υπηρεσίες, φωτογραφικές μηχανές
  - ✓Λίγες βαθμολογήσεις
  - ✓Περασμένων ετών
- Έχουν βαθιά γνώση του τομέα των προϊόντων
- Απουσία προβλήματος ψυχρής εκκίνησης
- Σαφώς δυσκολότερη υλοποίηση

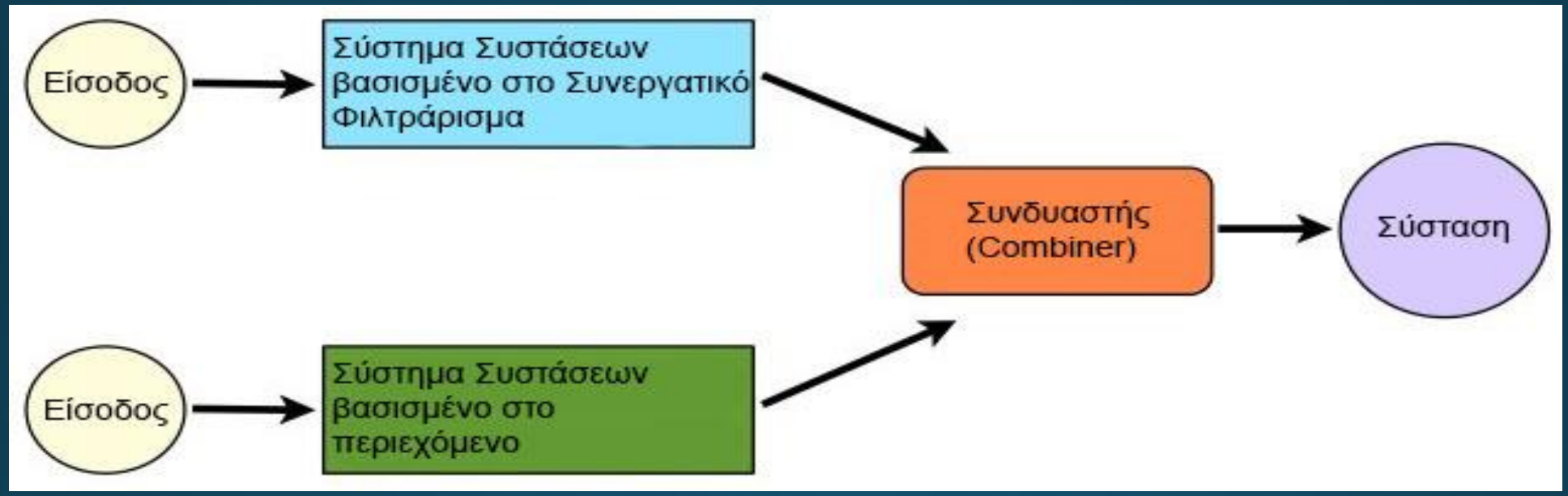
3)Βασισμένα στη Γνώση

- Ανατροφοδότηση των χρηστών μέσω ερωτήσεων ή κριτικών
- Είδη συστημάτων βασισμένων στη γνώση
  1. Βασισμένα στην περίπτωση
  2. Βασισμένα σε περιορισμούς

4)Βασισμένα σε Δημογραφικά Δεδομένα

- Δημογραφικά στοιχεία: ηλικία, φύλο, ενδιαφέροντα
- Λογική: άνθρωποι με κοινά ενδιαφέροντα, κοντινές ηλικίες και ίδιο φύλο είναι πολύ πιθανό να έχουν κοινές προτιμήσεις
- Μειονεκτήματα:
  1. Πολύ γενικές συστάσεις
  2. Δεν προσαρμόζονται στις αλλαγές ενδιαφέροντος
  3. Άρνηση εισαγωγής πραγματικών στοιχείων
  4. Μη συμπλήρωση ερωτηματολογίων
- Συνδυάζονται με άλλες προσεγγίσεις

5)Υβριδικά



- Συνδυάζει πολλαπλές τεχνικές μαζί για να επιτύχει κάποια συνέργεια μεταξύ τους
- Συνεργατικό φιλτράρισμα – φιλτράρισμα βάσει περιεχομένου είναι αποτελεσματικό σε πολλές περιπτώσεις
- Πιο ακριβείς συστάσεις από καθαρές προσεγγίσεις

5) Υβριδικά

- Ξεπερνούν προβλήματα όπως ψυχρή εκκίνηση και το πρόβλημα των ελαχίστων αναφορών

Κεφάλαιο 3

Σύστημα Συστάσεων Βασισμένο σε Χωροχρονική Πληροφορία

Χωροχρονική πληροφορία

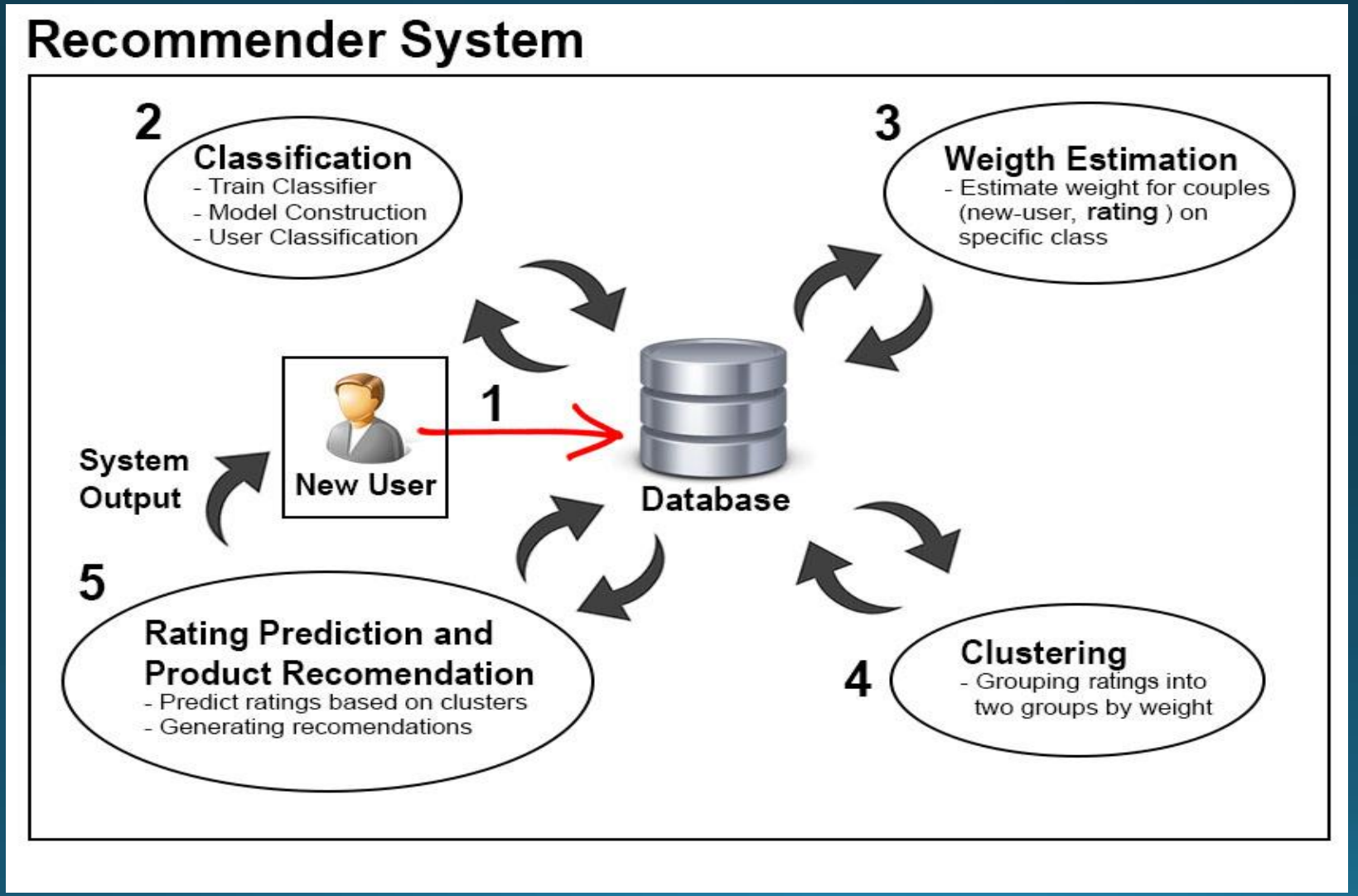
- Χρήστες: Γεωγραφική θέση στην οποία θα δεχτούν συστάσεις.

longitude	latitude	date_inserted
24.32275	40.921348	1998-04-23 01:10:38
23.022901	40.653807	1998-04-23 01:10:38
22.593822	39.728357	1998-04-23 01:10:38

- Βαθμολογίες: Γεωγραφική θέση και χρόνος που πραγματοποιήθηκαν.

id ▲ 1	user_id	item_id	rating	prediction	longitude	latitude	date_inserted
1	196	242	3	0	20.9793	38.0345	1997-12-04 16:55:49
2	186	302	3	0	20.9336	37.2433	1998-04-04 21:22:22
3	22	377	1	0	21.3451	39.8317	1997-11-07 08:18:36

Λειτουργία Συστήματος



Υποσυστήματα

- Το σύστημα αποτελείται από 4 υποσυστήματα

1. Κατηγοριοποίησης
2. Υπολογισμού βαρών
3. Συσταδοποίησης
4. Πρόβλεψης βαθμολογίας και εξαγωγής συστάσεων

Υποσυστήματα

- Κατηγοριοποίηση: Η διαδικασία κατά την οποία ο χρήστης τοποθετείται σε μία κατηγορία ανάλογα με δημογραφικά δεδομένα.
- Υπολογισμός βαρών: Η διαδικασία υπολογισμού ενός αριθμού που παραπέμπει στο βαθμό ομοιότητας των χαρακτηριστικών ενός χρήστη και μίας βαθμολογίας.
- Συσταδοποίηση: Η διαδικασία της ομαδοποίησης των βαθμολογιών με βάση τα βάρη τους.
- Πρόβλεψη βαθμολογίας και συστάσεις: Η διαδικασία πρόβλεψης βαθμολογίας με βάση όμοιες βαθμολογίες άλλων χρηστών και εξαγωγής συστάσεων από αυτές.

Κεφάλαιο 4

Κατηγοριοποίηση

Κατηγοριοποίηση

- Μέθοδος εξόρυξης δεδομένων κατά την οποία ένα στοιχείο ανατίθεται σε ένα προκαθορισμένο σύνολο κατηγοριών
- Διαδικασία 2 βημάτων:
  1. Κατασκευή του μοντέλου
  2. Αποτίμηση του μοντέλου

❑ Κατασκευή Μοντέλου:

- Σετ εκπαίδευσης
- Ανάλυση δεδομένων μέχρι να φτάσει σε ένα συμπέρασμα
- Κατηγοριοποίηση: Εποπτευόμενη μάθηση
- Διαφορετικοί αλγόριθμοι, διαφορετικές τεχνικές

Κατηγοριοποίηση

- Αποτίμηση του Μοντέλου
  - Δοκιμαστικά δεδομένα (test data), διαφορετικά από τα δεδομένα εκπαίδευσης
  - Βγάζει συμπέρασμα για την απόδοση του αλγορίθμου
  - Συγκρίνει την τιμή πρόβλεψης της κατηγορίας από τα δοκιμαστικά δεδομένα με την υπάρχουσα τιμή των δεδομένων εκπαίδευσης
- Είδη κατηγοριοποιητών:
  1. Δυαδικός κατηγοριοποιητής
  2. Κατηγοριοποιητής πολλαπλών κλάσεων

Αλγόριθμος Κατηγοριοποίησης

- Naïve Bayes
  - Τιμή ενός χαρακτηριστικού ανεξάρτητη της τιμής οποιαδήποτε άλλου χαρακτηριστικού
  - Χρησιμοποιεί τη μέθοδο μέγιστης πιθανοφάνειας. Επιλέγει την υπόθεση που είναι πιο πιθανή
  - Απαιτούν μικρό αριθμό δεδομένων για εκπαίδευση
  - Συνάρτηση που κάνει αντιστοίχιση των ετικετών:

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i|C_k).$$

Αλγόριθμος Κατηγοριοποίησης

- C4.5
  - Δημιουργία δέντρου απόφασης
  - Χρησιμοποιεί την έννοια των πληροφοριών εντροπίας

Υλοποίηση Κατηγοριοποίησης

- Κατηγοριοποίηση με βάση δημογραφικά δεδομένα
  1. Φύλο
  2. Ηλικία
- Πεπερασμένος αριθμός κλάσεων.
- Όνομα κλάσης: { Φύλο_δεκαετία }
- Παράδειγμα: Για ένα πελάτη που είναι 25 χρονών και άντρας η σωστή τιμή κλάσης είναι η man_twenty

birth_date	class	sex
1992-09-01	man_twenty	man
1963-10-26	woman_fifty	woman
1993-07-13	man_twenty	man
1992-11-05	man_twenty	man
1983-02-06	woman_thirty	woman
1974-07-25	man_fourty	man
1959-09-11	man_fifty	man
1980-03-26	man_thirty	man

Κεφάλαιο 5

Υπολογισμός Βαρών

Υπολογισμός βαρών

- Δεκαδικός αριθμός από το 0 έως το 1
- Ομοιότητα χαρακτηριστικών χρήστη – βαθμολογίας
- Μεγαλύτερο βάρος => Μεγαλύτερη ομοιότητα
- Βάρος για κάθε ζευγάρι πελάτη-βαθμολογίας (της οποίας ο βαθμολογητής έχει την ίδια τιμή κλάσης με τον πελάτη)
- Χαρακτηριστικά που εμπλέκονται στον υπολογισμό:
  1. Γεωγραφική θέση
  2. Χρόνος βαθμολόγησης
  3. Ηλικία βαθμολογητή

Συναρτήσεις Βαρών

- Τρόπος υπολογισμού του βάρους ζευγαριού { χρήστη-βαθμολογίας }

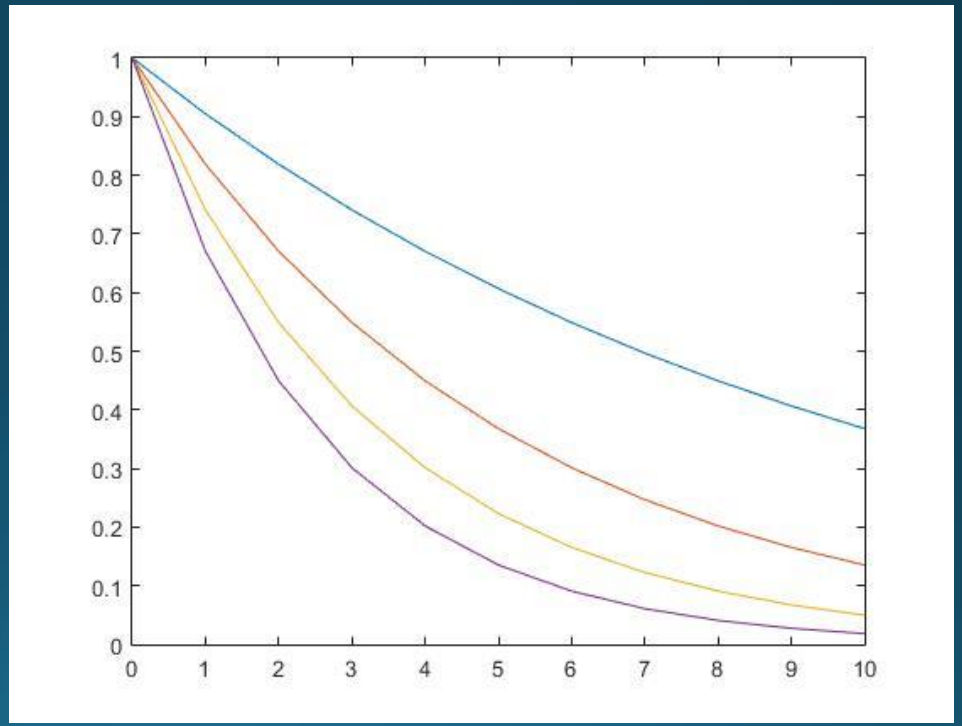
$$Weight = \kappa * ageSim + \lambda * timeSim + \mu * locSim$$
$$= \kappa * \frac{1}{e^{(\alpha * \Delta Years)}} + \lambda * \frac{1}{e^{(\gamma * \Delta time)}} + \mu * \frac{1}{e^{(\beta * \Delta Location)}}$$

- ageSim: συνάρτηση υπολογισμού βάρους ηλικίας βαθμολογητή
- timeSim: συνάρτηση υπολογισμού βάρους χρόνου βαθμολόγησης
- locSim: συνάρτηση υπολογισμού βάρους γεωγραφική θέσης
- Όπου κ, λ, μ δεκαδικοί αριθμοί από το 0 έως 1 που αντιπροσωπεύουν το ποσοστό με το οποίο συνεισφέρει κάθε χαρακτηριστικό στον υπολογισμό του συνολικού βάρους

Συναρτήσεις Βάρους Ηλικίας

- Όπου ΔYears η απόλυτη τιμή της διαφοράς της ηλικίας του πελάτη για τον οποίο προορίζονται οι συστάσεις από την ηλικία του βαθμολογητή.

$$ageSim = \frac{1}{e^{(\alpha * \Delta Years)}}$$

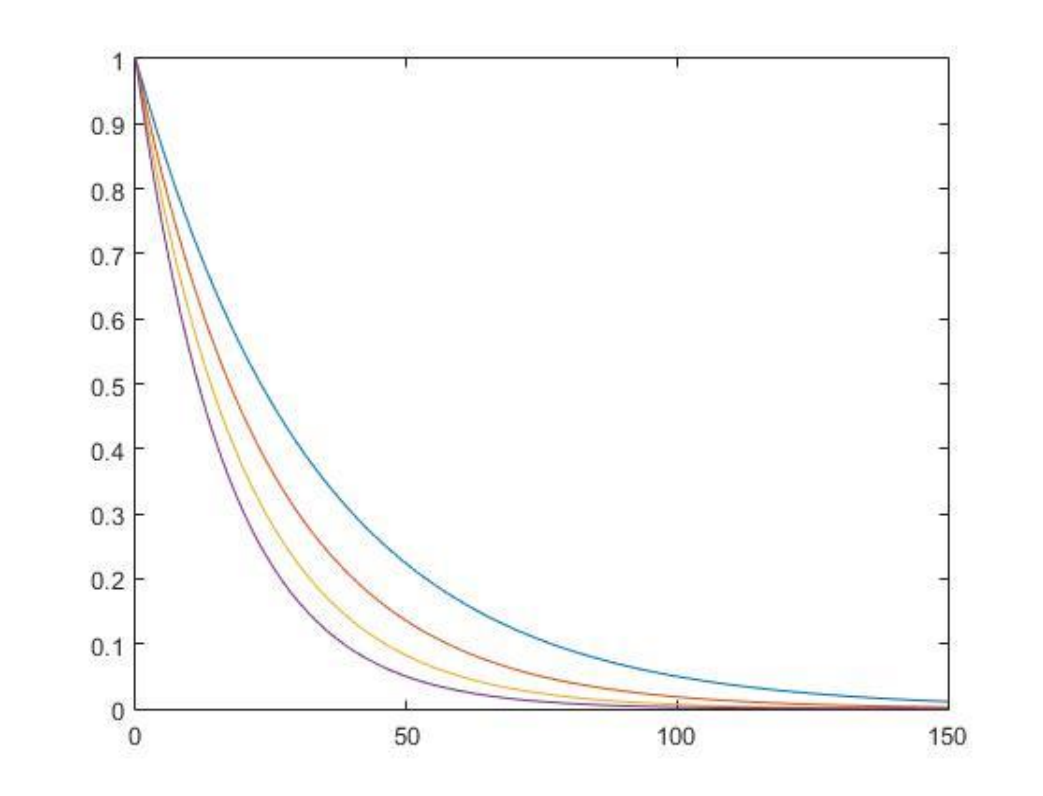


- Όσο μεγαλύτερο το α τόσο πιο απότομη η γραφική παράσταση της συνάρτησης

Συναρτήσεις Βάρους Γεωγραφικής Θέσης

- Όπου ΔLocation η απόλυτη τιμή της διαφοράς σε χιλιόμετρα της γεωγραφικής θέσης του πελάτη για τον οποίο προορίζονται οι συστάσεις από την αυτή του βαθμολογητή.

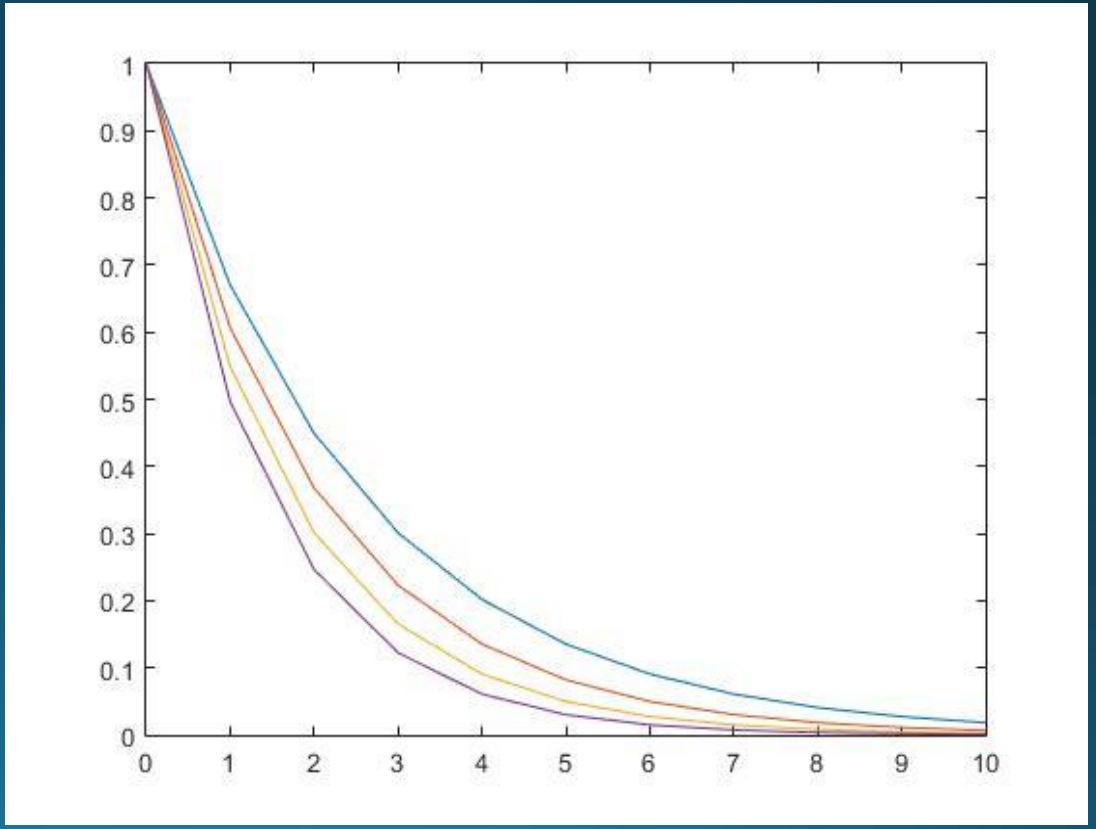
$$locSim = \frac{1}{e^{(\beta * \Delta Location)}}$$



Συναρτήσεις Βάρους Χρόνου

- Όπου ΔTime η απόλυτη τιμή της διαφοράς του χρόνου βαθμολόγησης από την ημερομηνία της σύστασης σε χρόνια.

$$timeSim = \frac{1}{e^{(\gamma * \Delta time)}}$$

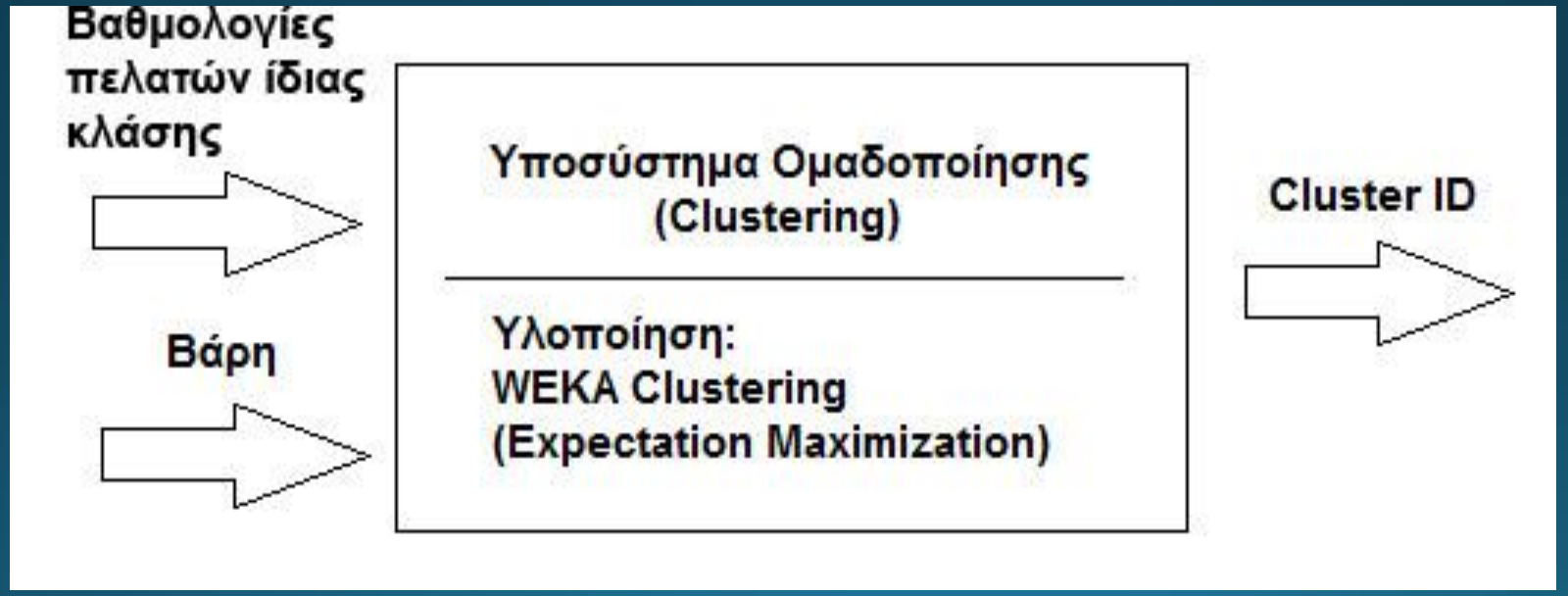


Κεφάλαιο 6

Συσταδοποίηση

Συσταδοποίηση

- Η Συσταδοποίηση είναι η διαδικασία της ομαδοποίησης ενός συνόλου αντικειμένων, με τέτοιο τρόπο ώστε όλα τα αντικείμενα μέσα στην ίδια ομάδα να είναι περισσότερο όμοια μεταξύ τους, σε σχέση με αυτά των άλλων ομάδων



Συσταδοποίηση

- Δημιουργία 2 συστάδων:
  1. Συστάδα υψηλού ενδιαφέροντος:
     - Σε αυτή ανήκουν οι βαθμολογίες με τα μεγαλύτερα βάρη.
     - Οι βαθμολογίες της διαμορφώνουν τις προβλέψεις βαθμολογίας
     - Προκύπτουν οι συστάσεις
  2. Συστάδα χαμηλού ενδιαφέροντος
     - Βαθμολογίες με μικρά βάρη
     - Πολύ μικρή πιθανότητα να ενδιαφέρουν το χρήστη
     - Αγνοούνται από το σύστημα

Αλγόριθμος Συσταδοποίησης (Expectation Maximization)

- Χωρίς στάδιο εκπαίδευσης
- Αλγόριθμος βελτιστοποίησης συνάρτησης
- Μεγιστοποιεί την αναμενόμενη τιμή της λογαριθμικής συνάρτησης πιθανοφάνειας δοθέντων των παρατηρούμενων δειγμάτων και της εκτίμησης θ της τρέχουσας επανάληψης.
- 2 Βήματα:
  - Βήμα αναμενόμενης τιμής: Υπολογισμός της αναμενόμενης τιμής της συνάρτησης λογαριθμικής πιθανοφάνειας.
  - Βήμα Μεγιστοποίησης: Υπολογισμός της επόμενης εκτίμησης της παραμέτρου θ (παράμετρος που θέλουμε να εκτιμήσουμε) μεγιστοποιώντας της λογαριθμικής συνάρτησης πιθανοφάνειας (μηδενίζοντας της παράγωγο ως προς θ και βρίσκοντας το μέγιστο).

Κεφάλαιο 7

Πρόβλεψη Βαθμολογίας και Εξαγωγή Συστάσεων

Πρόβλεψη βαθμολογίας

- Γενική εικόνα που έχουν οι χρήστες (γείτονες) για ένα προϊόν
- Δημιουργείται από την εξαγωγή των δεδομένων της συστάδας υψηλού ενδιαφέροντος
- Μέσος όρος των βαθμολογιών των προϊόντων
- Όριο για πρόβλεψη σύστασης: 10 βαθμολογίες
  - Μεγιστοποίηση της αξιοπιστίας των προβλέψεων

Εξαγωγή Συστάσεων

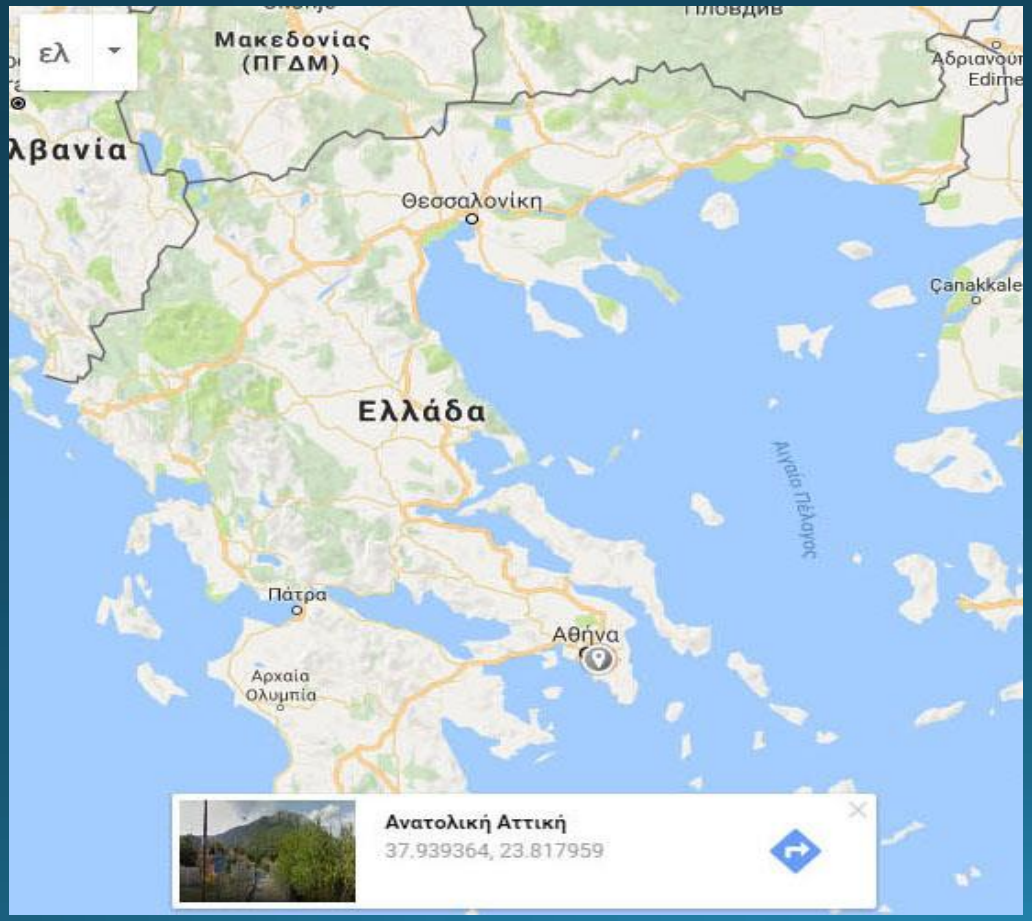
- Είσοδος: Προβλέψεις βαθμολογίας των προϊόντων που εμφανίζονται στη συστάδα υψηλού ενδιαφέροντος
- Έξοδος: Προτάσεις προϊόντων προς το χρήστη (συστάσεις)
- Μορφή: Λίστα 4 προϊόντων με τη μεγαλύτερη πρόβλεψη βαθμολογίας
  - Αρκετά μικρός για να διατηρηθεί η αξιοπιστία του συστήματος
  - Αρκετά μεγάλος ώστε να δίνει περισσότερες επιλογές σε περίπτωση αστοχίας κάποιας σύστασης

Κεφάλαιο 8

Πειραματική Αποτίμηση

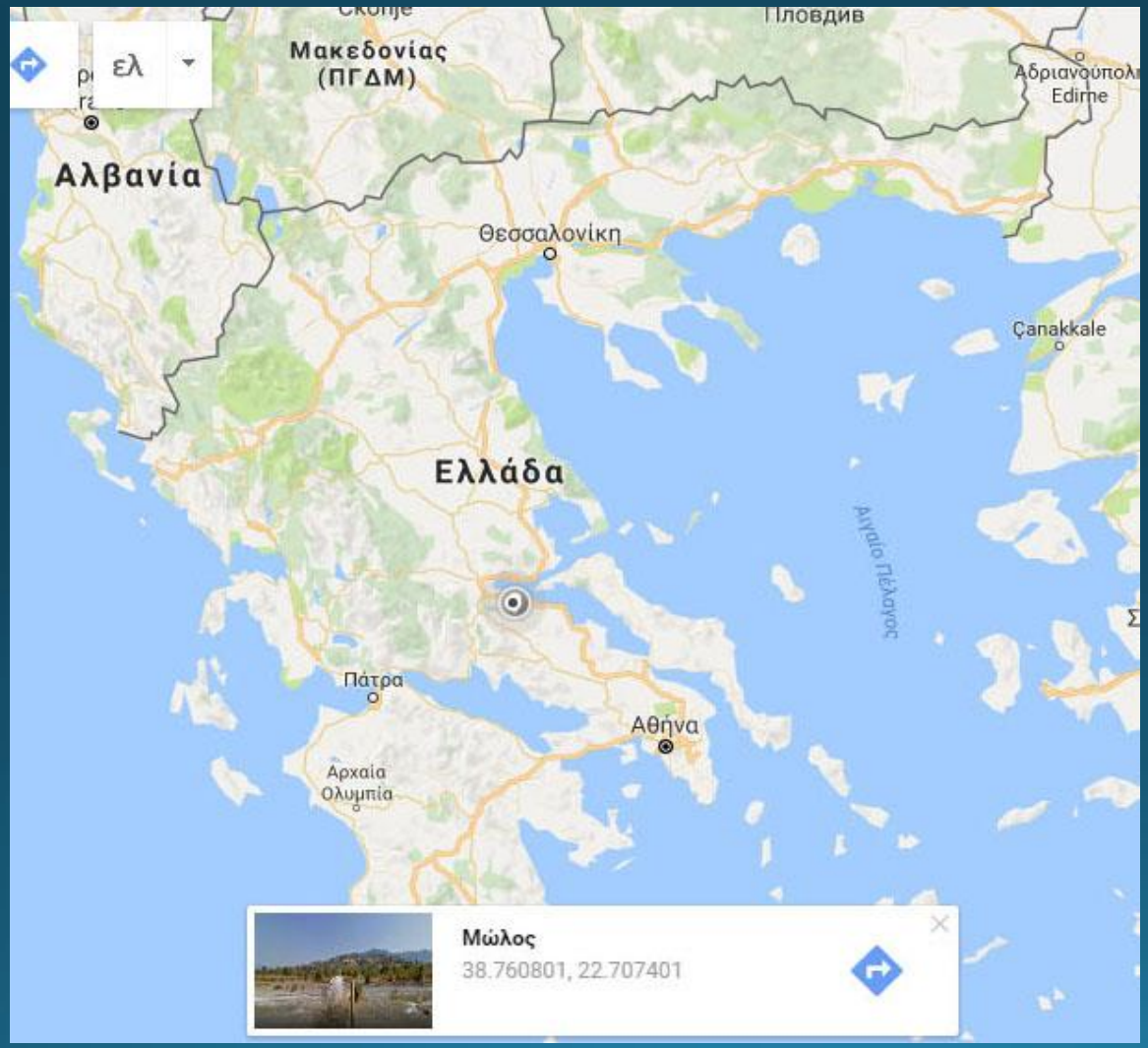
Πειραματική αποτίμηση

- Υποθέτουμε πως ο χρήστης ταξιδεύει και περνά από 5 θέσεις στις οποίες δέχεται συστάσεις
- Θέση Α:



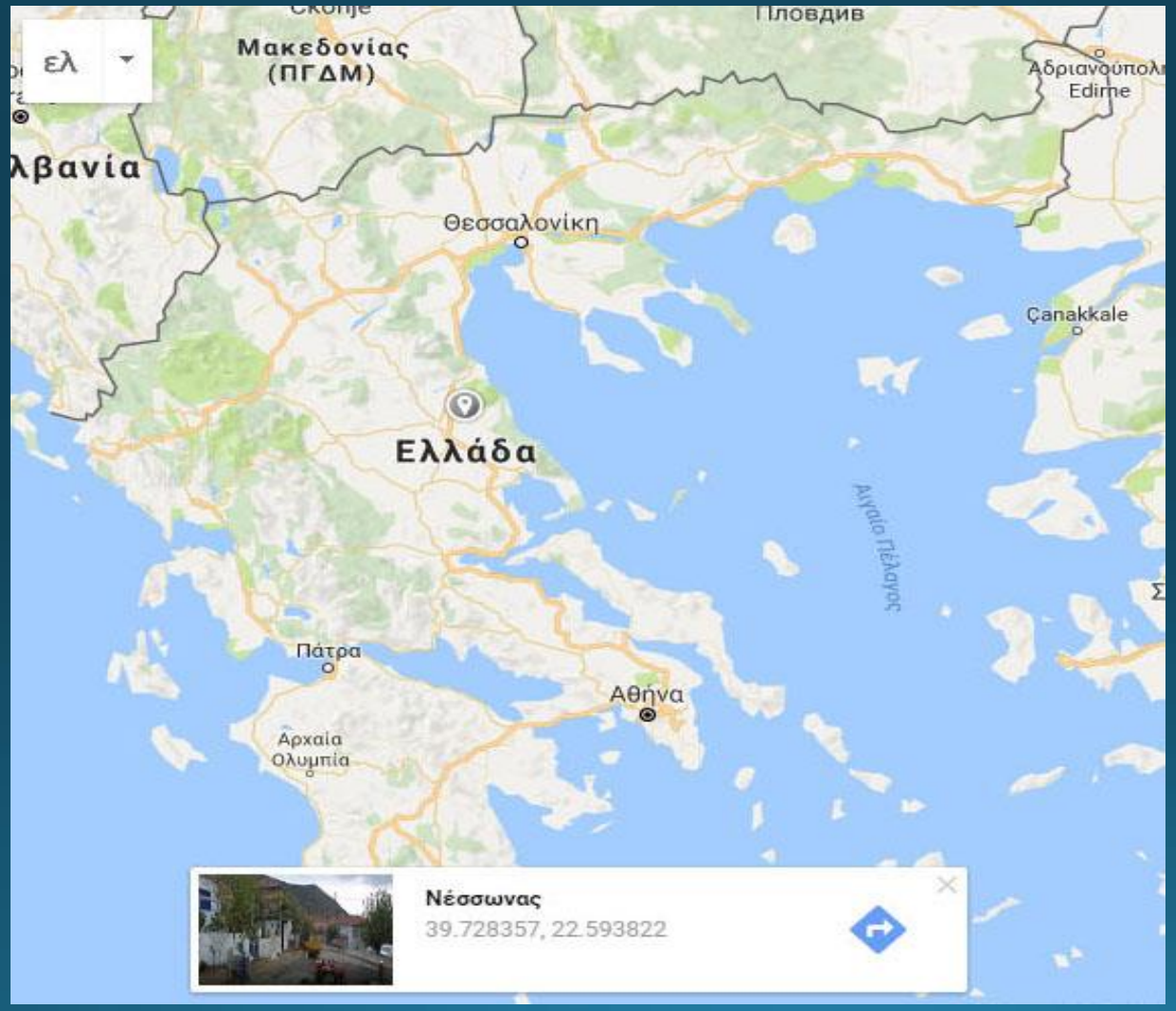
Πειραματική αποτίμηση

- Θέση Β:



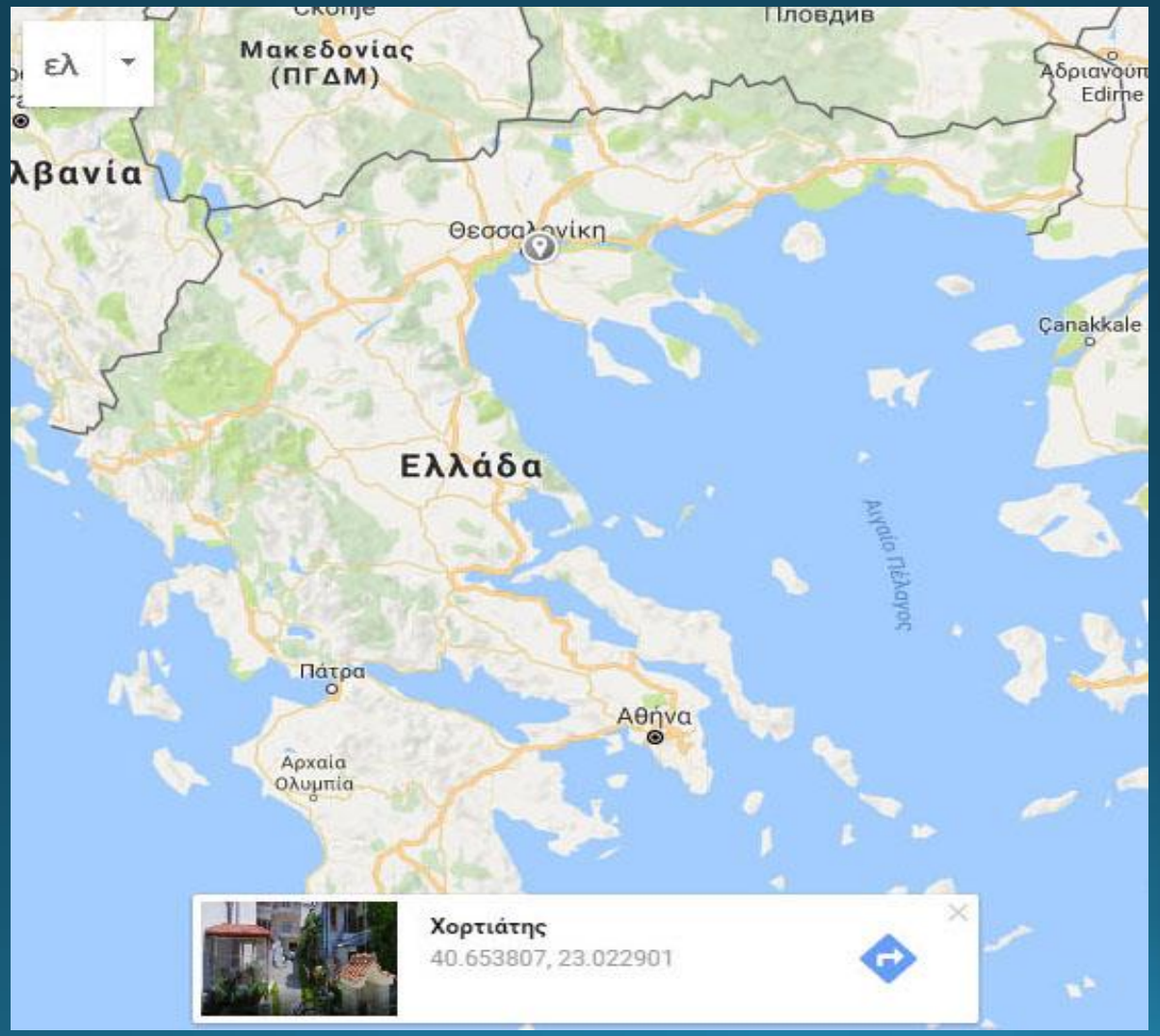
Πειραματική αποτίμηση

- Θέση Γ:



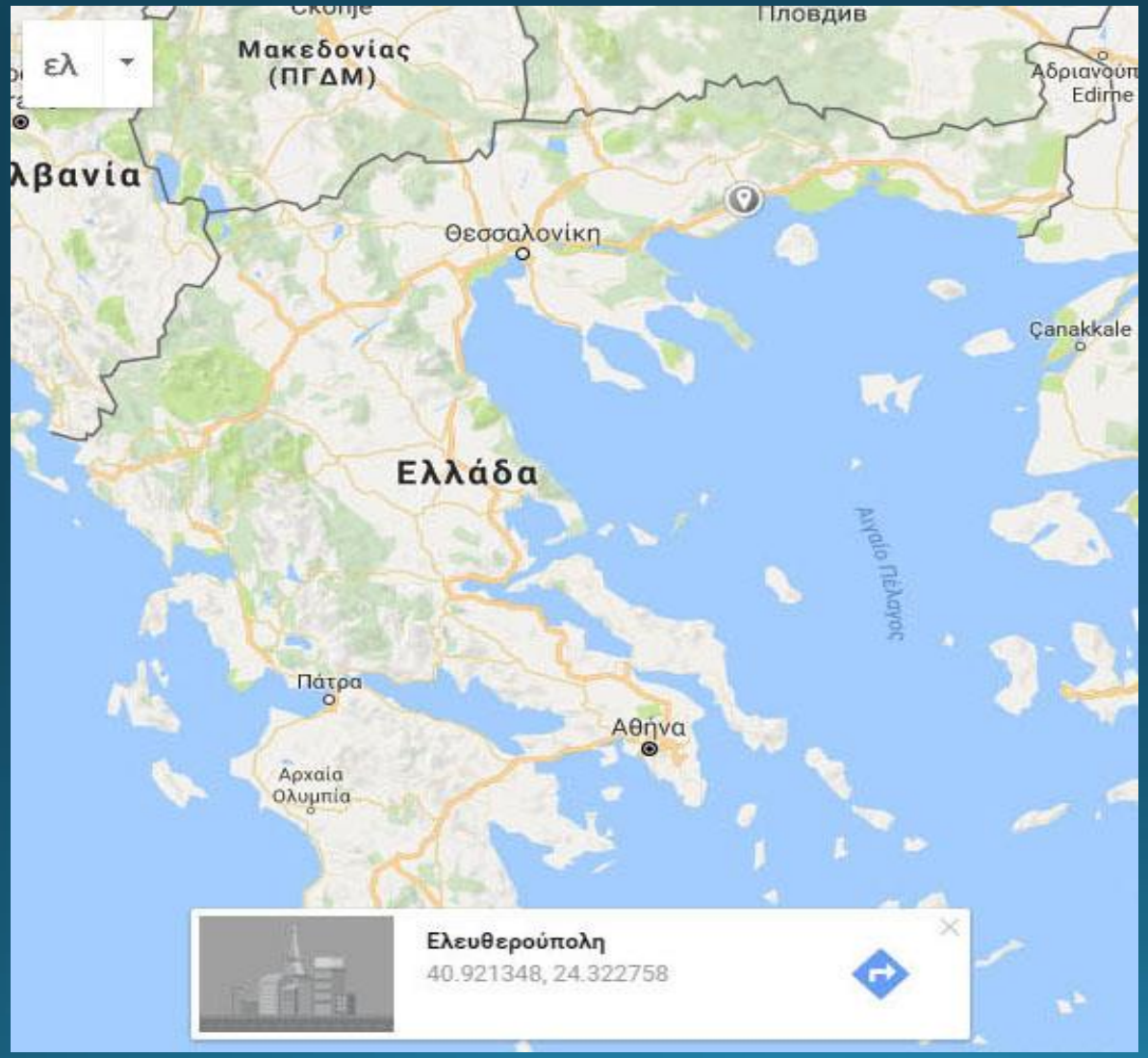
Πειραματική αποτίμηση

- Θέση Δ:



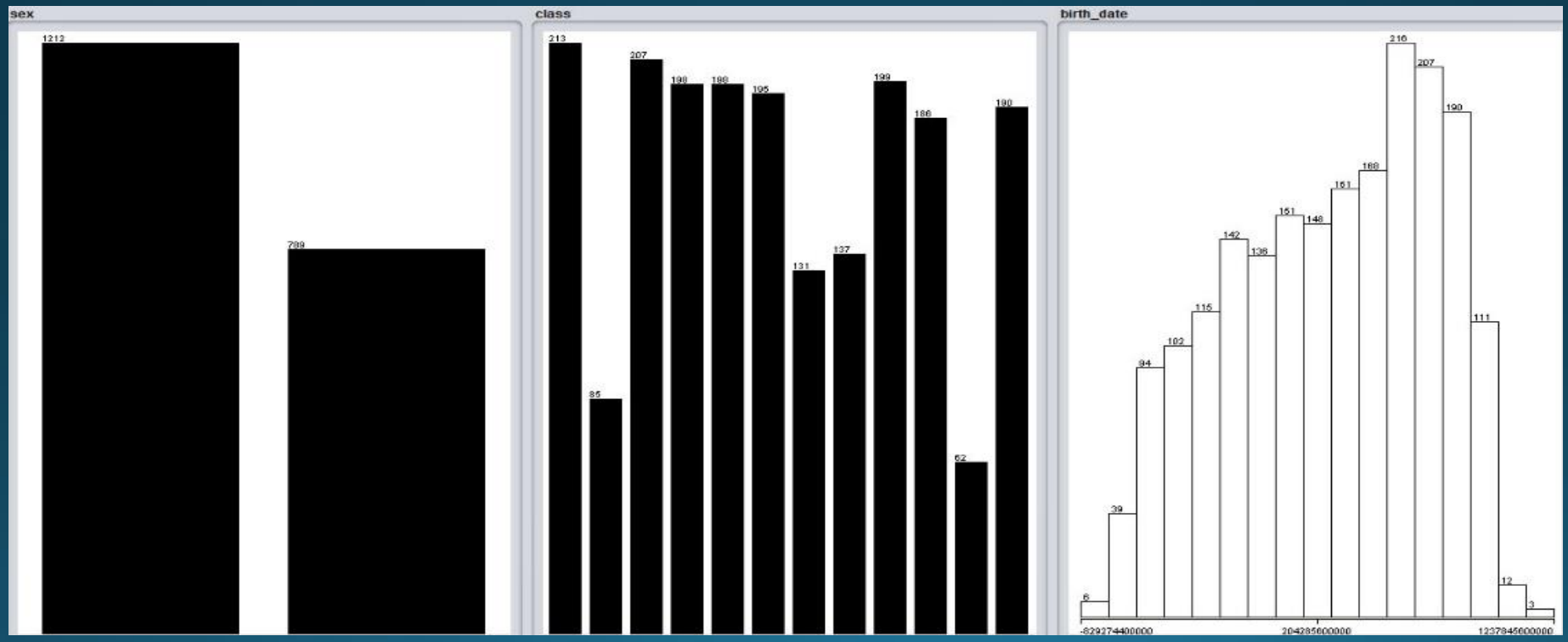
Πειραματική αποτίμηση

- Θέση Ε:



Αποτελέσματα Κατηγοριοποίησης (WEKA)

- Χρησιμοποίηση της εφαρμογής του WEKA



Naïve Bayes

- Εκπαίδευση του μοντέλου
  - 2001 χρήστες, 20% σετ εκπαίδευσης (training data)

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     QueryResult
Instances:    2001
Attributes:   3
              sex
              class
              decate
Test mode:    split 20.0% train, remainder test
```

  - Χαρακτηριστικά:
    1. Φύλο
    2. Κλάση
    3. Ηλικία

Naïve Bayes

- Δημιουργία μοντέλου

Naive Bayes Classifier

Attribute	Class 2 (0.27)	5 (0.16)	3 (0.21)	4 (0.19)	0 (0)	1 (0.08)	6 (0.09)	7 (0)
sex								
man	351.0	183.0	270.0	218.0	2.0	96.0	96.0	4.0
woman	183.0	136.0	157.0	165.0	1.0	70.0	84.0	1.0
[total]	534.0	319.0	427.0	383.0	3.0	166.0	180.0	5.0
class								
man_twenty	351.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
woman_fifty	1.0	136.0	1.0	1.0	1.0	1.0	1.0	1.0
woman_thirty	1.0	1.0	157.0	1.0	1.0	1.0	1.0	1.0
man_fourty	1.0	1.0	1.0	218.0	1.0	1.0	1.0	1.0
man_fifty	1.0	183.0	1.0	1.0	1.0	1.0	1.0	1.0
man_thirty	1.0	1.0	270.0	1.0	1.0	1.0	1.0	1.0
woman_twenty	183.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
woman_fourty	1.0	1.0	1.0	165.0	1.0	1.0	1.0	1.0
man_zero	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0
woman_ten	1.0	1.0	1.0	1.0	1.0	70.0	1.0	1.0
man_ten	1.0	1.0	1.0	1.0	1.0	96.0	1.0	1.0
man_sixty	1.0	1.0	1.0	1.0	1.0	1.0	96.0	1.0
woman_sixty	1.0	1.0	1.0	1.0	1.0	1.0	84.0	1.0
man_seventy	1.0	1.0	1.0	1.0	1.0	1.0	1.0	4.0
[total]	546.0	331.0	439.0	395.0	15.0	178.0	192.0	17.0

- Υποθέτει ότι οι κλάσεις είναι ισοπίθανες. Για ν κλάσεις $p(x_1)=p(x_2)=\ldots=p(x_n)$
- Αποτέλεσμα $p(x_1|y)>p(x_2|y)$ => κλάση x1 ΑΛΛΙΩΣ κλάση x2

Naïve Bayes

- Αποτίμηση μοντέλου:
  - Από 2001 πελάτες, 80% test data
  - Χρόνος:

```
=== Evaluation on test split ===

Time taken to test model on training split: 0.01 seconds
```

  - Αποτελέσματα:

```
=== Summary ===

Correctly Classified Instances        1599               99.8751 %
Incorrectly Classified Instances         2                0.1249 %
Kappa statistic                          0.9985
Mean absolute error                      0.0348
Root mean squared error                  0.0616
Relative absolute error                 17.1698 %
Root relative squared error             19.3581 %
Total Number of Instances             1601
```

  - 99,88% επιτυχία! Στις 1601 κατηγοριοποιήσεις 1599 σωστές και 2 λάθος!

Naïve Bayes

- Confusion Matrix:

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   <-- classified as
 428   0   0   0   0   0   0   0 |   a = 2
   0 259   0   0   0   0   0   0 |   b = 5
   0   0 335   0   0   0   0   0 |   c = 3
   0   0   0 303   0   0   0   0 |   d = 4
   0   0   0   0   0   0   0   0 |   e = 0
   0   0   0   0   0 125   0   0 |   f = 1
   0   0   0   0   0   0 149   0 |   g = 6
   0   0   2   0   0   0   0   0 |   h = 7
```

C4.5

- Εκπαίδευση του μοντέλου
  - 2001 χρήστες, 20% σετ εκπαίδευσης (training data)
  - Χαρακτηριστικά:
    1. Φύλο
    2. Κλάση
    3. Ηλικία

C4.5

- Εκπαίδευση του μοντέλου
  - 2001 χρήστες, 20% σετ εκπαίδευσης (training data)
  - Χαρακτηριστικά:
    1. Φύλο
    2. Κλάση
    3. Ηλικία

C4.5

- Δημιουργία του μοντέλου

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

class = man_twenty: 2 (350.0)
class = woman_fifty: 5 (135.0)
class = woman_thirty: 3 (156.0)
class = man_fourty: 4 (217.0)
class = man_fifty: 5 (182.0)
class = man_thirty: 3 (269.0)
class = woman_twenty: 2 (182.0)
class = woman_fourty: 4 (164.0)
class = man_zero: 0 (1.0)
class = woman_ten: 1 (69.0)
class = man_ten: 1 (95.0)
class = man_sixty: 6 (95.0)
class = woman_sixty: 6 (83.0)
class = man_seventy: 7 (3.0)

Number of Leaves  :     14

Size of the tree :      15
```

- Χτίζει δέντρα απόφασης από ένα σύνολο δεδομένων εκπαίδευσης, χρησιμοποιώντας την έννοια της εντροπίας πληροφοριών
- Κάθε φύλλο του δέντρου είναι μία ξεχωριστή κλάση

C4.5

- Αποτίμηση μοντέλου:
  - Από 2001 πελάτες, 80% test data
  - Αποτελέσματα:

```
=== Summary ===

Correctly Classified Instances        1601               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             1601
```

  - 100% επιτυχία! Ακόμα καλύτερος από τον Naïve Bayes

C4.5

- Confusion Matrix

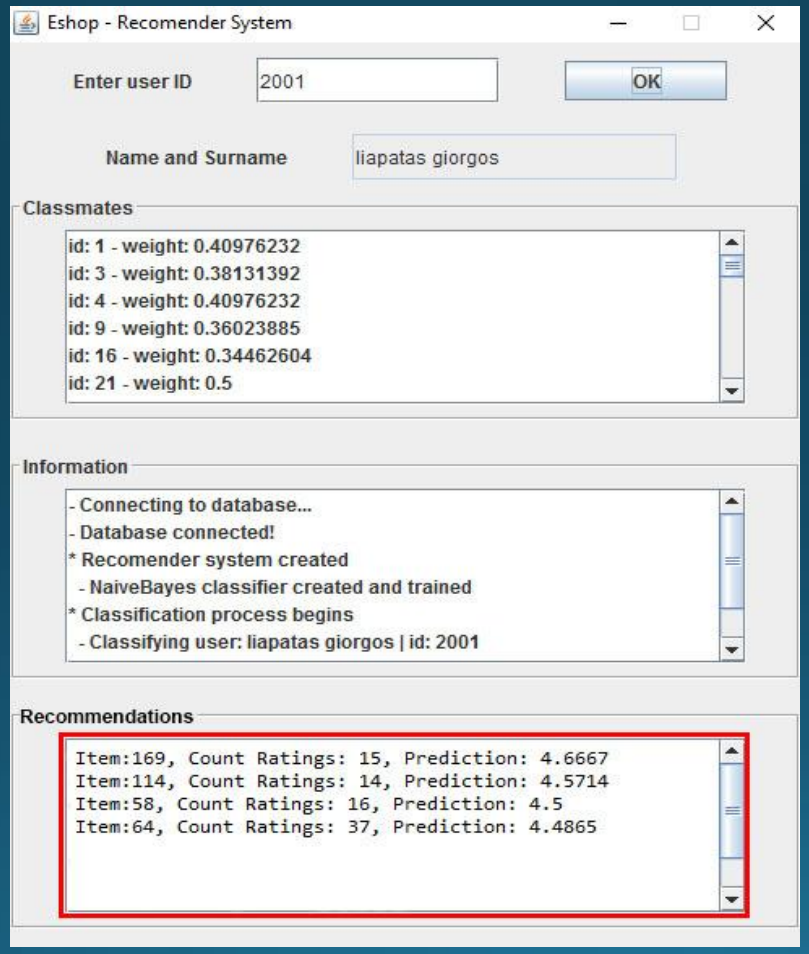
```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   <-- classified as
 428   0   0   0   0   0   0   0 |   a = 2
   0 259   0   0   0   0   0   0 |   b = 5
   0   0 335   0   0   0   0   0 |   c = 3
   0   0   0 303   0   0   0   0 |   d = 4
   0   0   0   0   0   0   0   0 |   e = 0
   0   0   0   0   0 125   0   0 |   f = 1
   0   0   0   0   0   0 149   0 |   g = 6
   0   0   0   0   0   0   0   2 |   h = 7
```

Αλγόριθμος Expectation-Maximization

- Δε χρειάζεται εκπαίδευση

```
=== Run information ===

Scheme:       weka.clusterers.EM -I 100 -N 2 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation:     QueryResult
Instances:    198
Attributes:   2
              classmate_id
              weight
Test mode:    evaluate on training data
```

Μορφή αποτελεσμάτων



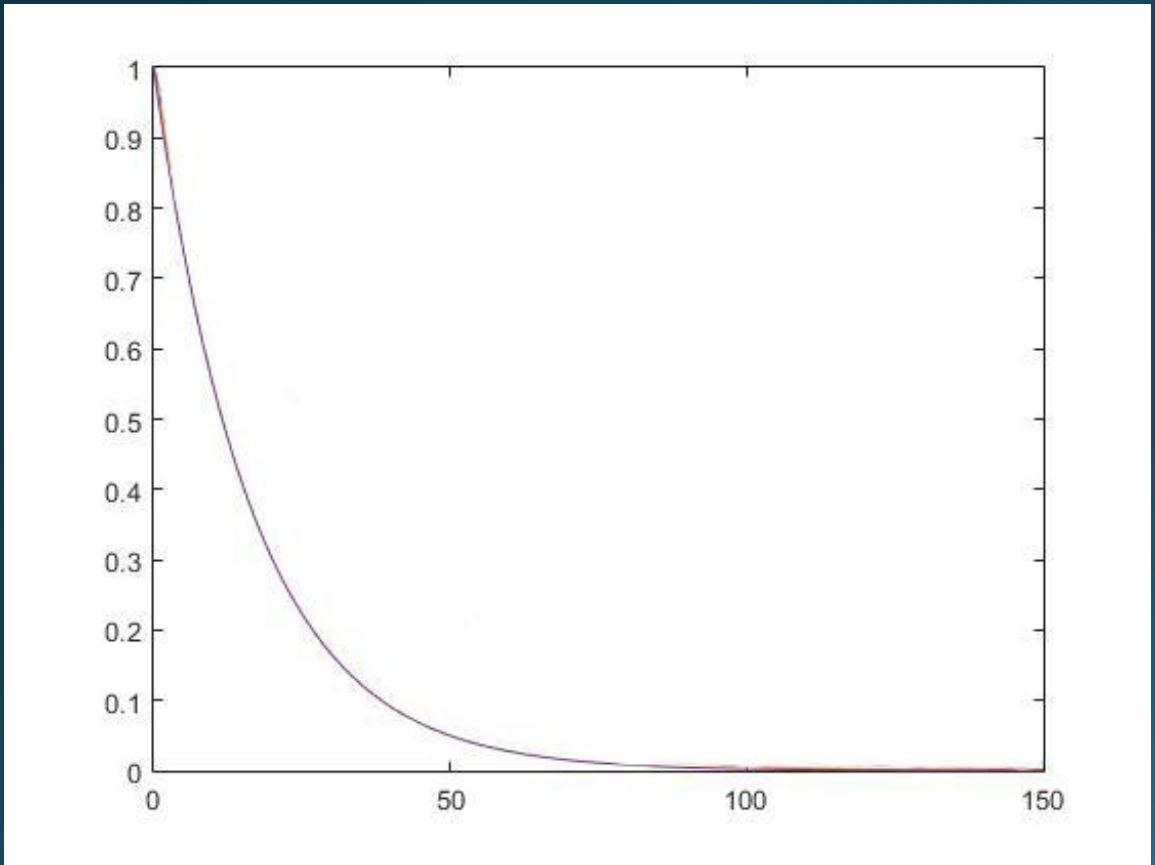
Πειραματική αποτίμηση 3 σεναρίων

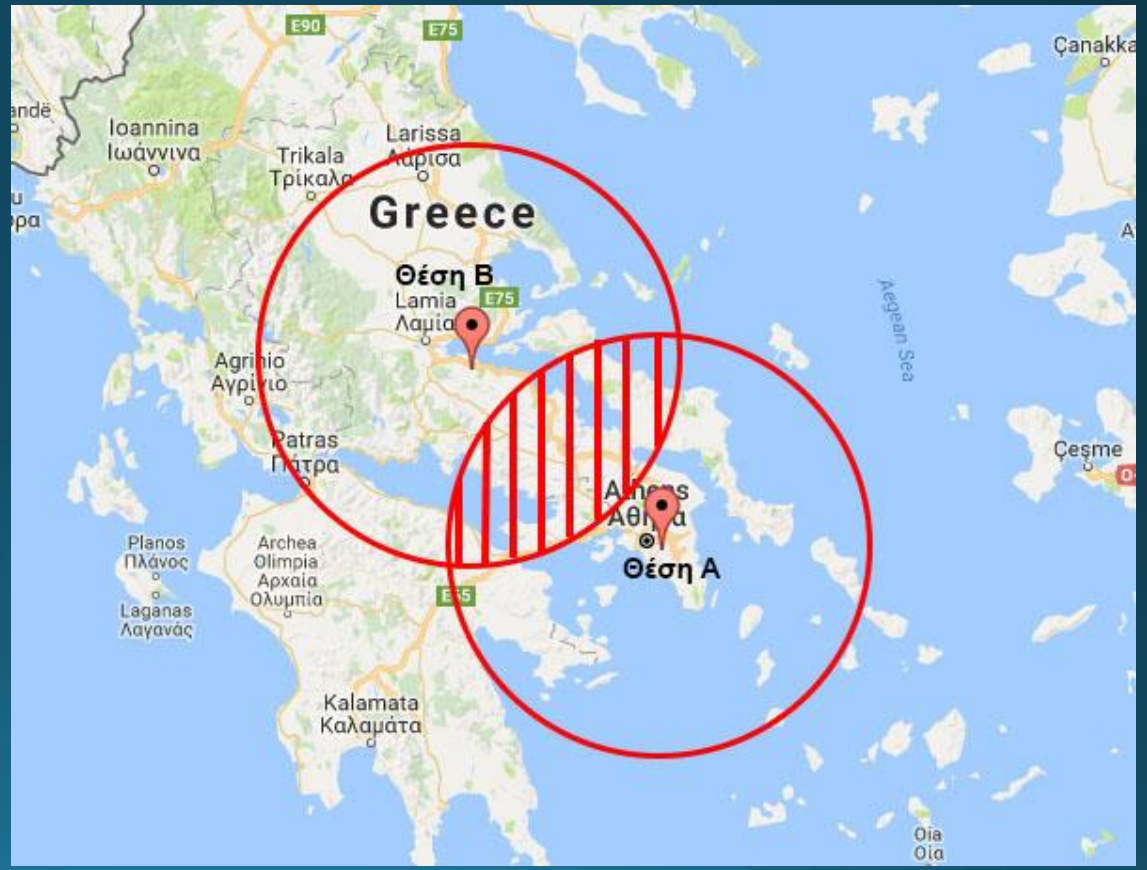
- Αποτελέσματα σεναρίου Α (**μ=1, λ=0, κ=0**)

```
Θέση Α
Rating Prediction => Item:50, Count Ratings: 17, Prediction: 4.8235
Rating Prediction => Item:98, Count Ratings: 12, Prediction: 4.5833
Rating Prediction => Item:79, Count Ratings: 11, Prediction: 4.5455
Rating Prediction => Item:174, Count Ratings: 14, Prediction: 4.4286

Θέση Β
Rating Prediction => Item:127, Count Ratings: 13, Prediction: 4.6154
Rating Prediction => Item:174, Count Ratings: 15, Prediction: 4.6
Rating Prediction => Item:50, Count Ratings: 14, Prediction: 4.5714
Rating Prediction => Item:79, Count Ratings: 12, Prediction: 4.25

Θέση Γ
Rating Prediction => Item:98, Count Ratings: 12, Prediction: 4.3333
Rating Prediction => Item:174, Count Ratings: 18, Prediction: 4.2778
Rating Prediction => Item:50, Count Ratings: 13, Prediction: 4.1538
Rating Prediction => Item:172, Count Ratings: 16, Prediction: 4.125


Θέση Δ
Rating Prediction => Item:150, Count Ratings: 13, Prediction: 4.4615
Rating Prediction => Item:96, Count Ratings: 15, Prediction: 4.4
Rating Prediction => Item:50, Count Ratings: 17, Prediction: 4.2941
Rating Prediction => Item:181, Count Ratings: 12, Prediction: 4.25

Θέση Ε
Rating Prediction => Item:313, Count Ratings: 11, Prediction: 4.5455
Rating Prediction => Item:50, Count Ratings: 13, Prediction: 4.3846
Rating Prediction => Item:176, Count Ratings: 12, Prediction: 4.25
Rating Prediction => Item:237, Count Ratings: 12, Prediction: 4.1667
```

Σενάριο Α (μ=1, λ=0, κ=0)

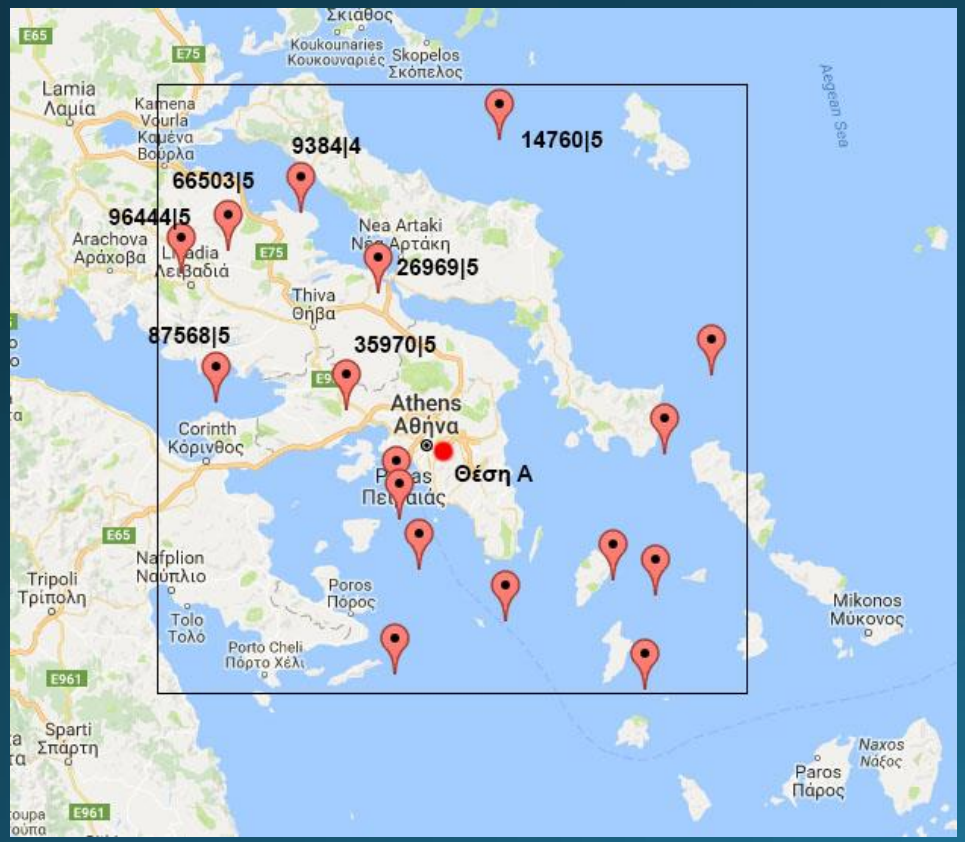
- Αλληλοκάλυψη περιοχών που δημιουργεί η συνάρτηση locSim

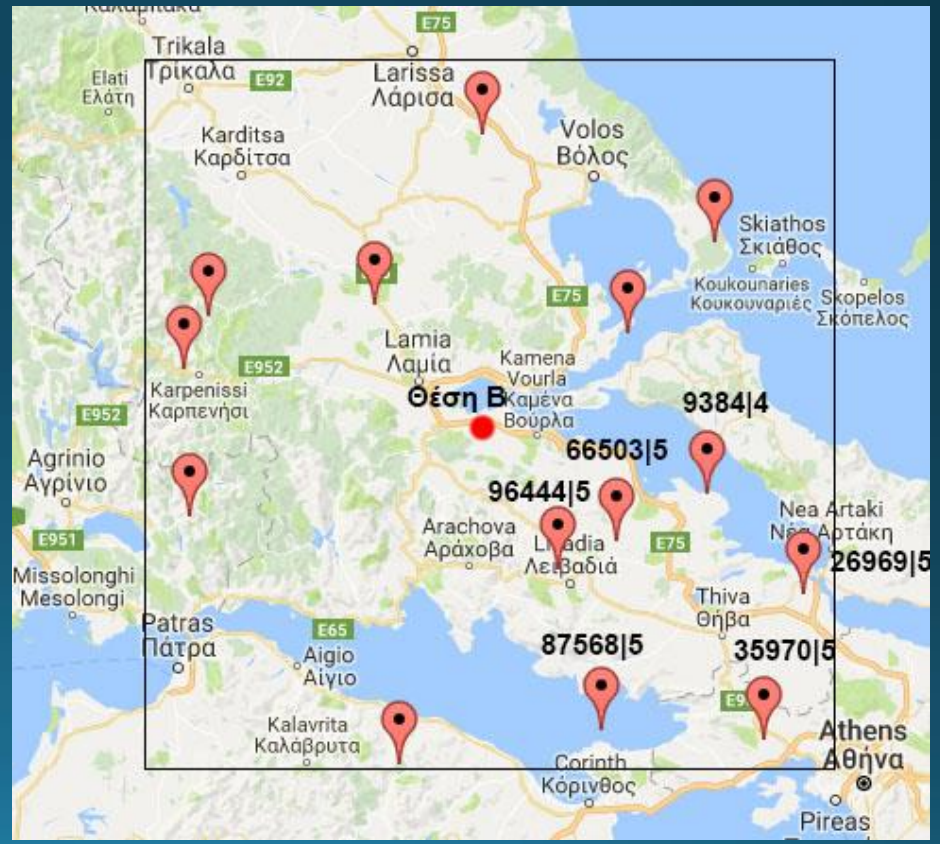




Σενάριο Α (μ=1, λ=0, κ=0)

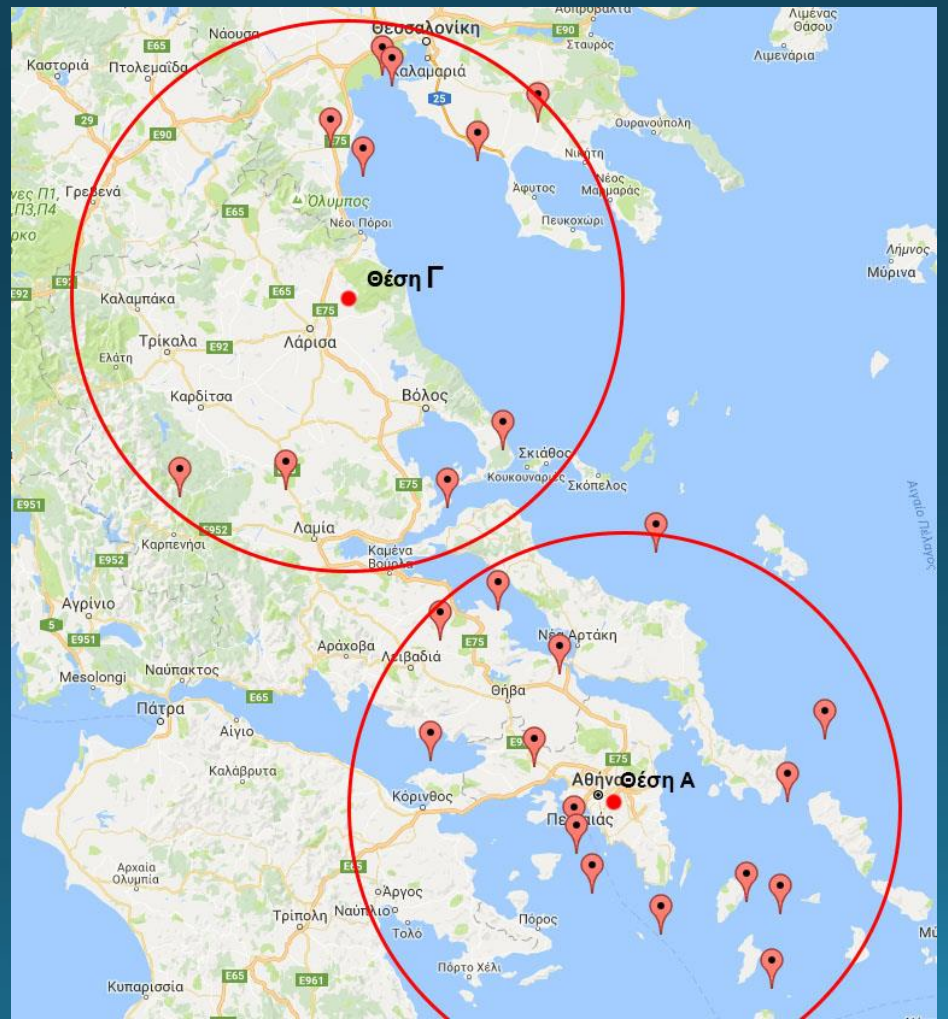
- Αλληλοκάλυψη περιοχών (Θέση Α και Θέση Β) για το προϊόν με ID 50.





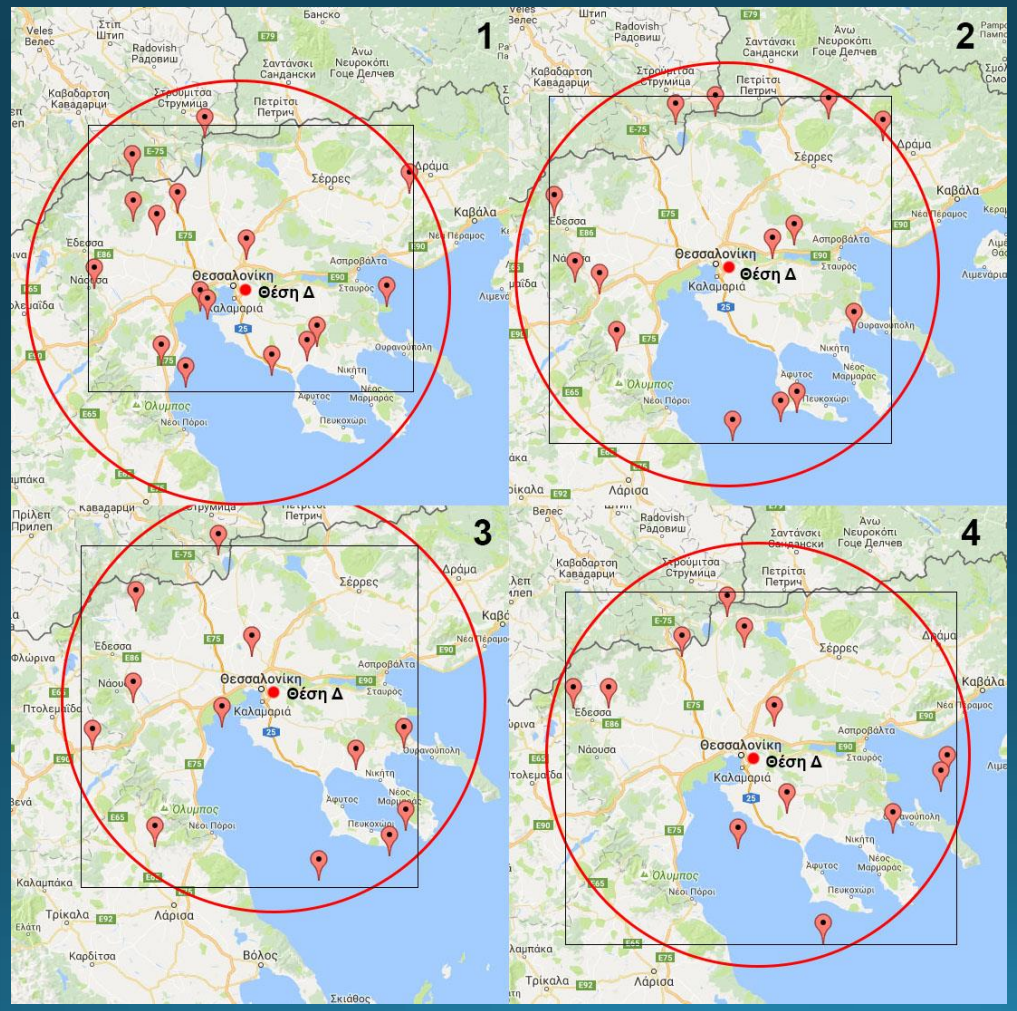
Σενάριο Α (μ=1, λ=0, κ=0)

- Βαθμολογίες προϊόντος με ID 50 για τις θέσεις Α και Γ.



Σενάριο Α (μ=1, λ=0, κ=0)

- Βαθμολογίες που διαμορφώνουν τις συστάσεις της θέσης Δ



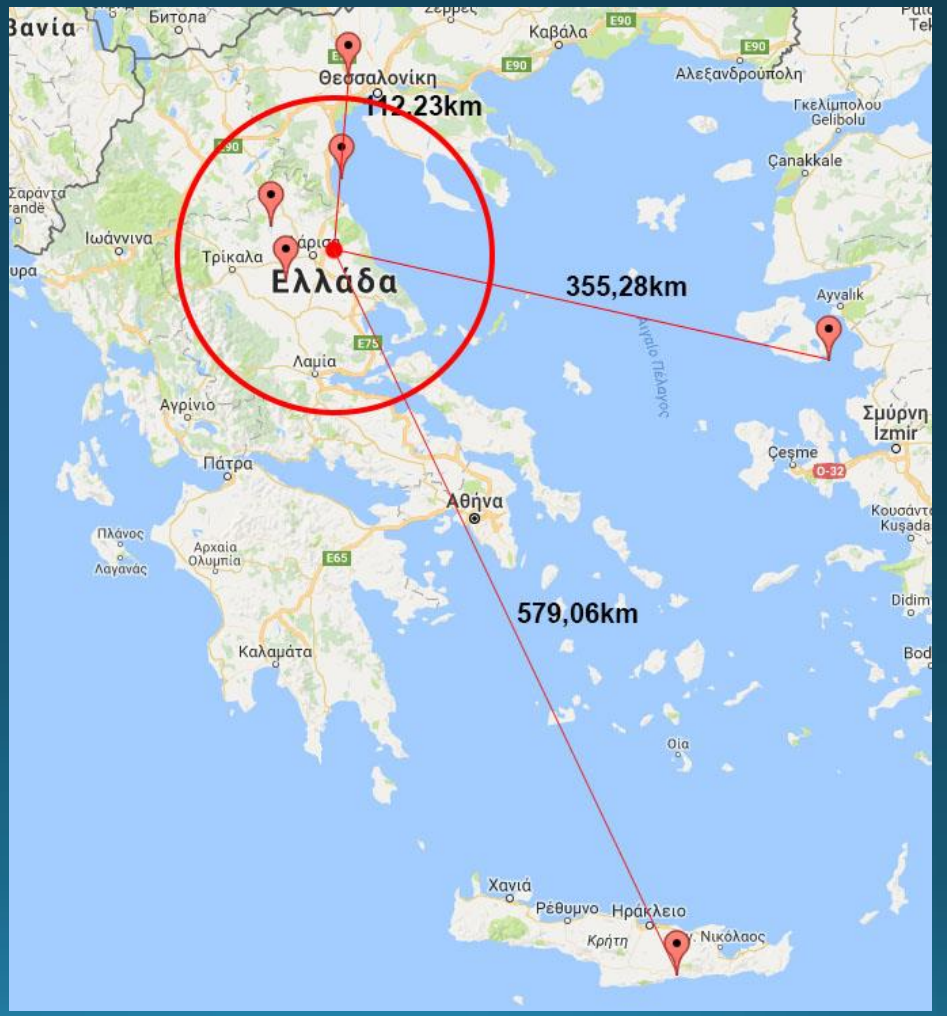
Πειραματική αποτίμηση 3 σεναρίων

- Αποτελέσματα σεναρίου Β (**μ=0.8, λ=0.1, κ=0.1**)

```
Θέση Α
Rating Prediction => Item:50, Count Ratings: 7, Prediction: 4.8571
Rating Prediction => Item:258, Count Ratings: 6, Prediction: 4.1667
Rating Prediction => Item:7, Count Ratings: 6, Prediction: 3.5
Rating Prediction => Item:117, Count Ratings: 7, Prediction: 3.4286

Θέση Β
Rating Prediction => Item:50, Count Ratings: 7, Prediction: 4.5714
Rating Prediction => Item:100, Count Ratings: 9, Prediction: 4.2222
Rating Prediction => Item:288, Count Ratings: 8, Prediction: 4.0
Rating Prediction => Item:268, Count Ratings: 6, Prediction: 4.0

Θέση Γ
Rating Prediction => Item:174, Count Ratings: 6, Prediction: 4.8333
Rating Prediction => Item:96, Count Ratings: 6, Prediction: 4.3333
Rating Prediction => Item:313, Count Ratings: 6, Prediction: 4.1667
Rating Prediction => Item:288, Count Ratings: 6, Prediction: 4.1667

Θέση Δ
Rating Prediction => Item:50, Count Ratings: 7, Prediction: 4.1429
Rating Prediction => Item:313, Count Ratings: 7, Prediction: 4.1429
Rating Prediction => Item:181, Count Ratings: 8, Prediction: 4.125
Rating Prediction => Item:315, Count Ratings: 6, Prediction: 4.0

Θέση Ε
Rating Prediction => Item:176, Count Ratings: 6, Prediction: 4.5
Rating Prediction => Item:313, Count Ratings: 10, Prediction: 4.4
Rating Prediction => Item:1, Count Ratings: 6, Prediction: 3.8333
Rating Prediction => Item:100, Count Ratings: 8, Prediction: 3.125
```

Σενάριο Β (μ=0.8, λ=0.1, κ=0.1)

- Βαθμολογίες που διαμορφώνουν την πρόβλεψη βαθμολογίας του προϊόντος 288 για τη θέση Γ:



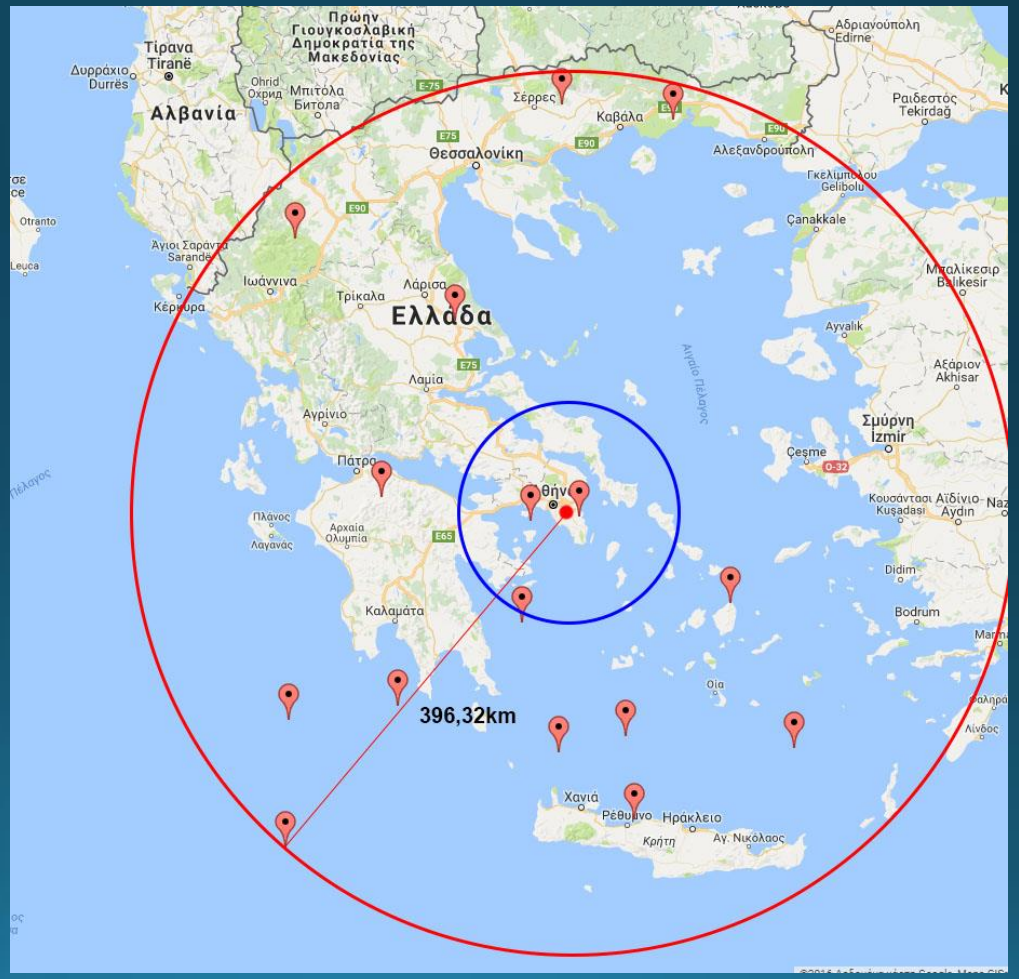
Πειραματική αποτίμηση 3 σεναρίων

- Αποτελέσματα σεναρίου Γ (**μ=0.5, λ=0.3, κ=0.2**)

```
Θέση Α
Rating Prediction => Item:174, Count Ratings: 16, Prediction: 4.5625
Rating Prediction => Item:318, Count Ratings: 8, Prediction: 4.5
Rating Prediction => Item:192, Count Ratings: 6, Prediction: 4.5
Rating Prediction => Item:211, Count Ratings: 6, Prediction: 4.5

Θέση Β
Rating Prediction => Item:174, Count Ratings: 14, Prediction: 4.5714
Rating Prediction => Item:518, Count Ratings: 6, Prediction: 4.5
Rating Prediction => Item:211, Count Ratings: 6, Prediction: 4.5
Rating Prediction => Item:318, Count Ratings: 9, Prediction: 4.4444

Θέση Γ
Rating Prediction => Item:174, Count Ratings: 15, Prediction: 4.6
Rating Prediction => Item:318, Count Ratings: 8, Prediction: 4.5
Rating Prediction => Item:22, Count Ratings: 10, Prediction: 4.4
Rating Prediction => Item:12, Count Ratings: 8, Prediction: 4.375

Θέση Δ
Rating Prediction => Item:135, Count Ratings: 6, Prediction: 4.6667
Rating Prediction => Item:174, Count Ratings: 14, Prediction: 4.5714
Rating Prediction => Item:318, Count Ratings: 8, Prediction: 4.5
Rating Prediction => Item:242, Count Ratings: 6, Prediction: 4.5

Θέση Ε
Rating Prediction => Item:174, Count Ratings: 14, Prediction: 4.5714
Rating Prediction => Item:318, Count Ratings: 8, Prediction: 4.5
Rating Prediction => Item:22, Count Ratings: 11, Prediction: 4.3636
Rating Prediction => Item:12, Count Ratings: 11, Prediction: 4.3636
```

Σενάριο Γ (μ=0.5, λ=0.3, κ=0.2)

- Βαθμολογίες που διαμορφώνουν την πρόβλεψη βαθμολογίας του προϊόντος 174 στη θέση Α



Συμπέρασμα

- Όσο μειώνεται το ποσοστό της συμβολής της γεωγραφικής θέσης, τόσο μεγαλύτερο είναι το ποσοστό των βαθμολογιών που βρίσκονται μακρύτερα από το χρήστη.
- Η σύσταση χάνει τον τοπικό χαρακτήρα της.

Αριθμητική αποτίμηση

- Μετρικές απόδοσης:
  - Απόλυτο Σφάλμα:

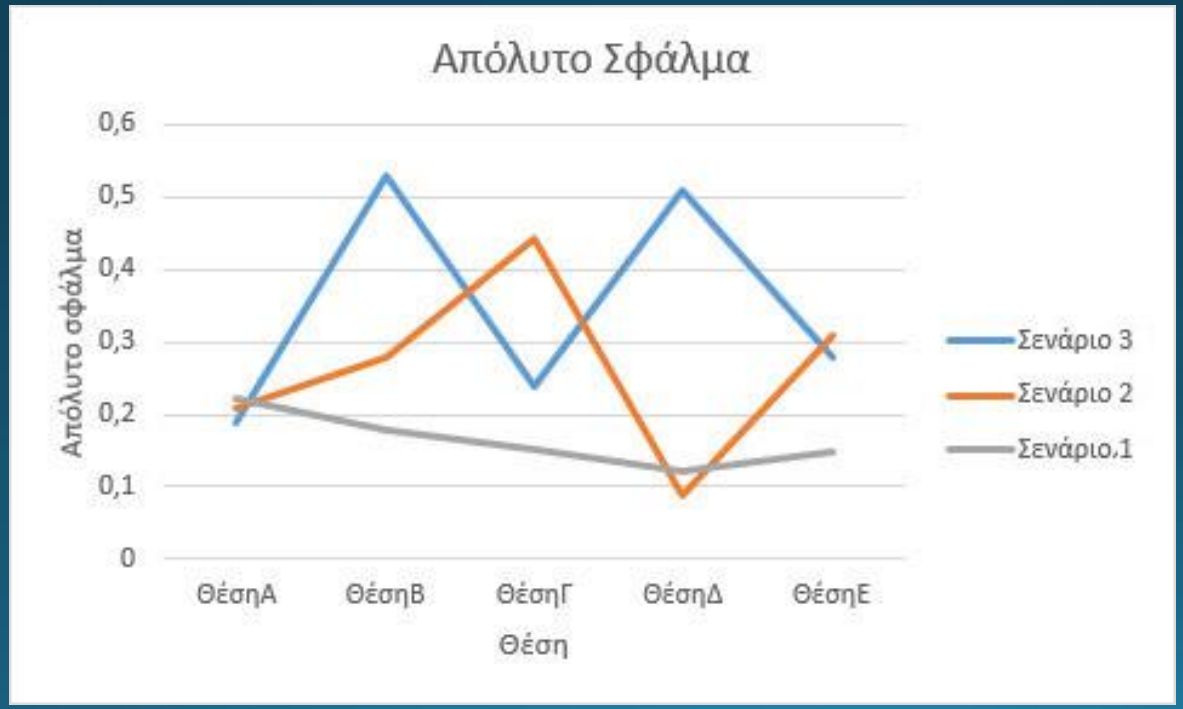
$$MAE = \frac{1}{N}\sum_{ij} |prediction_{ij} - real_{ij}|$$

  - Μέσο Τετραγωνικό Σφάλμα:

$$RMSE = \sqrt{\frac{1}{N}\sum_{ij} (prediction_{ij} - real_{ij})^2}$$

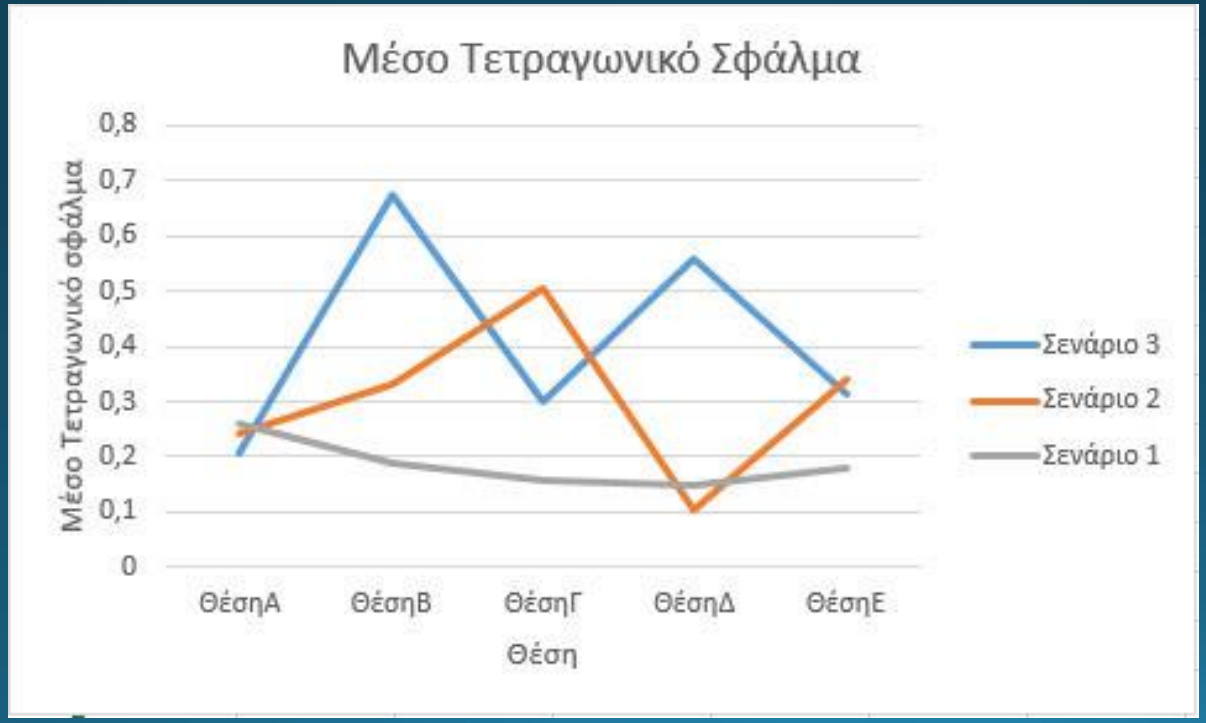
Απόλυτο Σφάλμα

	Θέση Α	Θέση Β	Θέση Γ	Θέση Δ	Θέση Ε
Σενάριο 1	0.2226	0.1797	0.15	0.1231	0.14675
Σενάριο 2	0.2091	0.28	0.4421	0.087	0.3079
Σενάριο 3	0.188	0.53	0.24	0.51	0.28



Μέσο Τετραγωνικό Σφάλμα

	Θέση Α	Θέση Β	Θέση Γ	Θέση Δ	Θέση Ε
Σενάριο 1	0.26	0.1887	0,1576	0.1465	0.1787
Σενάριο 2	0.2401	0.33	0.5066	0.1055	0.34
Σενάριο 3	0.205	0.6753	0.3	0.56	0.312



Κεφάλαιο 9

Συμπεράσματα και Μελλοντικές Προεκτάσεις

Συμπεράσματα

- Διαφορετική συμπεριφορά του συστήματος ανάλογα με το μίγμα χαρακτηριστικών που χρησιμοποιήθηκε για τον υπολογισμού των βαρών κάθε βαθμολογίας.
- Μείωση ποσοστού συμβολής της γεωγραφικής θέσης στο συνολικό βάρος:
  1. Αύξηση του σφάλματος
  2. Απώλεια τοπικού χαρακτήρα συστάσεων => καθολικές συστάσεις
  3. Επανάληψη των ίδιων βαθμολογιών στη διαμόρφωση των προβλέψεων

Μελλοντικές Προεκτάσεις

- Αποδοχή συστάσεων και ανατροφοδότηση
- Εμπλοκή μεγαλύτερου αριθμού δημογραφικών χαρακτηριστικών όπως είναι το επάγγελμα στην κατηγοριοποίηση των χρηστών
  - Πιο στοχευμένες συστάσεις

Ερωτήσεις